Smooth Post-Stratification for Multiple Capture-Recapture

By Zachary T. Kurtz

Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, U.S.A.
zkurtz@stat.cmu.edu

Summary

When the approximate size of a population is of primary interest and the cost of an exhaustive enumeration is prohibitive, capture-recapture models can be used to estimate the size of the population using only incomplete lists. Log-linear models that express the probability of each capture pattern in terms of the relative sizes of list-intersections tend to be biased by heterogeneity of capture probabilities across population units. We introduce a smooth generalization of post-stratification that allows the full suite of log-linear modeling tools to be applied at the unit level for closed populations. We illustrate the generality and simplicity of our novel approach by estimating bird species richness in continental North America.

Some key words: Capture-recapture; Continuous Covariate; Post-stratification; Closed Population; Species Richness.

1. Introduction

In many statistical applications, it is important to know the total size of a population when only samples are available; that is, to estimate how much of the population was not seen. Capture-recapture methods attempt to estimate population size from multiple samples and their patterns of overlap, an undertaking which requires careful modeling of the sampling process. Populations studied using capture-recapture are diverse, including various animal species (Odum & Pontin, 1961; Pollock et al., 1984), human populations (Chen et al., 2010), and the set of errors in a body of computer code (Runeson & Wohlin, 1998). This paper gives a new approach to the underlying statistical problem of estimating the size of a population from multiple incomplete lists or samples.

We review some basic capture-recapture concepts before introducing our new approach. In the simplest setting, there are two lists. Assume that units can be perfectly matched across lists, so it is possible to cross-classify units by list membership as in Table 1. Here c_{ij} is the count of units

Table 1. *A two-list cross-classification array*

		List 2	
		yes	no
List 1	yes	c_{11}	c_{10}
	no	c_{01}	$c_{00} =?$

with capture pattern ij. For example, c_{10} is the number of units on List 1 but not on List 2. The unknown number of units that are not observed on either list is denoted c_{00}, and estimating the population size amounts to estimating c_{00}. With three lists, the task is to estimate c_{000}, and so on.

The Petersen estimator is $\hat{c}_{00} = c_{10}c_{01}/c_{11}$, which can be formalized as a maximum-likelihood estimator under certain assumptions (Pollock, 1976). Perhaps the strongest of these assumptions is that the lists are independent; the event that a unit is captured on the first list is independent of the event that a unit is captured on the second list. Dependence between lists has two sources. The first source is unit-level list dependence, such as respondent fatigue, in which previous capture directly reduces the probability of subsequent capture. The second source of dependence arises indirectly as a consequence of heterogeneity, or variability in capture probabilities across units (Fienberg et al., 1999). In particular, when the capture probabilities are positively (negatively) correlated across lists, the Petersen estimator tends to be biased downwards (upwards).

Both sources of dependence may vary with covariates such as age. Post-stratification is an old method of accounting for covariates by partitioning the observed units into a finite collection of post-strata. For example, in a human population, observed individuals may be classified by age as younger than 30, between 30 and 60, or older than 60. Based on these post-strata, three separate classification arrays (as in Table 1) may be analyzed separately, resulting in estimates for the three missing cells, which can be summed to get an estimate of the total number of unobserved units. Post-stratification allows models that rely on homogeneity to perform reasonably well if each of the chosen post-strata is relatively homogeneous.

Because post-stratification has obvious disadvantages, including the loss of information in the discretization of continuous covariates, several modern models allow both unit-level dependence and heterogeneity to smoothly depend on covariates. Huggins (1989) and Alho (1990) developed logistic regression models for capture probability heterogeneity in the two-list scenario, and Yip et al. (2001) extended logistic regression to k lists with a simple respondent fatigue effect. A rather general semi-parametric regression model was developed by Hwang & Huggins (2011). Several other approaches are closely related: Chen & Lloyd (2002), Zwane & van der Heijden (2004), and Stoklosa & Huggins (2012).

We present a smooth generalization of post-stratification with two stages. The first stage is estimating the conditional probability of each capture pattern as a function of the covariates. The second stage is imputing the conditional probability of the unobserved capture pattern (no captures) by fitting a separate log-linear model for each observed population unit. The log-linear models imply an estimate of the number of unobserved units corresponding to each observed unit. Although some existing approaches mentioned above are similar, our framework is especially general, admitting both parametric and nonparametric estimation of the conditional probability of the capture pattern. More importantly, we allow log-linear model selection to vary over the covariate space. Throughout, we assume that the population is closed, precluding the possibility of births, deaths, and migration.

2. METHODS

2·1. *Notation and general framework*

Suppose k lists $L_1, ..., L_k$ are drawn from a population of unknown size n. Let $i = 1, ..., n_c$ index the units that are on at least one list. For each unit i and list L_j, let $y_{ij} = I(i \in L_j)$ be the indicator that the ith population unit appears on the jth list. Then $y_{i\cdot} = (y_{i1}, ..., y_{ik})$, and $y_{\cdot\cdot}$ is the $n \times k$ matrix with ith row $y_{i\cdot}$. The vector $y_{i\cdot}$ is called the capture pattern of the ith unit. Let $x_{i\cdot}$ denote a $1 \times q$ vector of covariates associated with the ith unit, and $x_{\cdot\cdot}$ is the $n \times q$ matrix with ith row $x_{i\cdot}$. For each $i > n_c$, the pair $(x_{i\cdot}, y_{i\cdot})$ is not observed. If $x_{\cdot\cdot}^c$ is the matrix formed by the first n_c rows of $x_{\cdot\cdot}$, and $y_{\cdot\cdot}^c$ is the matrix formed by the first n_c rows of $y_{\cdot\cdot}$, then the observable data is the pair of matrices $(x_{\cdot\cdot}^c, y_{\cdot\cdot}^c)$. We refer to the pair $(x_{\cdot\cdot}, y_{\cdot\cdot})$ as the extended data.

Let $\mathcal{Y}_k$ denote the set of binary row vectors of length k, so each $y_{i\cdot}$ is an element of $\mathcal{Y}_k$. Let $c_y := |\{i : y_{i\cdot} = y\}|$. The array $c := \{c_y : y \in \mathcal{Y}_k\}$ is the contingency table of counts of units in the lists. In particular, let $c_0 := c_{\vec{0}} = n - n_c$, the number of units that are not observed on any list. Assume that $y_{i\cdot}$ is a realization of a random vector $Y_{i\cdot}$. Let $p(i, y) = pr(Y_{i\cdot} = y)$, the probability that unit i has capture pattern y. Then $p(i, y_{i\cdot}) = pr(Y_{i\cdot} = y_{i\cdot})$.

We state a key assumption that is necessary for auxiliary-covariate models of heterogeneity that has often been left unstated in the literature. Namely, we assume that a smooth function $r(y, x)$ exists such that $p(i, y_{i\cdot}) = r(y_{i\cdot}, x_{i\cdot})$ $(i = 1, ..., n)$. This is a rather strong assumption, requiring that the covariates x fully explain any variation in the capture probabilities.

Let $\vec{0}^T$ denote the zero vector of length k. Define the detection function $\psi(x) = 1 - r(\vec{0}, x)$, which is the probability that a unit with covariates x appears in at least one of the lists. The Horvitz-Thompson estimator of the population size n can be written as

$$\tilde{n} = \sum_{i=1}^{n_c} \frac{1}{\psi(x_{i\cdot})}. \tag{1}$$

The estimator $\tilde{n}$ relies on the detection probabilities for only the units that are observed. To use (1), we must estimate the detection function ψ. If ψ is known, then $\tilde{n}$ has some nice asymptotic properties. It is easy to verify that $E(\tilde{n}) = n$. Moreover, $\tilde{n}$ is consistent and asymptotically normal if $\psi(x_{i\cdot})$ is uniformly bounded away from 0 and 1 (Alho, 1990).

The Horvitz-Thompson estimator (1) has been applied using various estimators for the detection function ψ. We propose a general way of framing these estimators. Define a function

$$\pi(y, x) := \frac{r(y, x)}{\sum_{z \neq \vec{0}} r(z, x)} = \frac{r(y, x)}{\psi(x)}. \tag{2}$$

For each nonzero $y \in \mathcal{Y}_k$, the function $\pi(y, x)$ is the conditional probability that a unit with covariates x will have capture pattern y, given that the unit is observed on at least one list. Conditioning on observation means that these functions are estimable directly from the data using any kind of binary regression. Holding y fixed, consider $\pi(y, x)$ as a function of x, and gather these functions into an array $\Pi := \{\pi(y, x) : y \neq \vec{0}\}$.

Rearranging (2) as $r(y, x) = \psi(x)\pi(y, x)$, it is easy to see that

$$\psi(x) = \frac{\sum_{y \neq \vec{0}} r(y, x)}{\sum_y r(y, x)} = \frac{\sum_{y \neq \vec{0}} \pi(y, x)}{\pi(\vec{0}, x) + \sum_{y \neq \vec{0}} \pi(y, x)} = \frac{1}{\pi(\vec{0}, x) + 1} \tag{3}$$

Equation (3) makes it seem natural to break the process of estimating ψ into two stages. The first stage is to generate estimates $\hat{\Pi} := \{\hat{\pi}(y, x) : y \neq \vec{0}\}$, while the second stage is to impute an estimator $\hat{\pi}(\vec{0}, x)$ from $\hat{\Pi}$. We deal with these two in sections 3 and 4 respectively.

Plugging the final expression from (3) into (1) leads to estimators involving the sum of the unit-level imputations:

$$\hat{n} := n_c + \sum_{i=1}^{n_c} \hat{\pi}(\vec{0}, x_{i\cdot}), \ \hat{c}_0 := \hat{n} - n_c = \sum_{i=1}^{n_c} \hat{\pi}(\vec{0}, x_{i\cdot}). \tag{4}$$

Any estimator $\widehat{\Pi}$ is a post-stratifier, and any smooth estimator $\widehat{\Pi}$ is a smooth post-stratifier. Finally, a smooth post-stratification estimator takes the form of (4), where $\hat{\pi}(\vec{0}, x)$ is imputed from a smooth estimator $\widehat{\Pi}$ (see Section 4).

Although it has not been stated in such an explicit and general form, smooth post-stratification is not new. An early version was derived under the title of local post-stratification (Chen &

Lloyd, 2002). We advocate the term smooth rather than local because post-stratification is already (i.e., redundantly) an attempt at localization; the relevant feature of the new framework is the assumption of a smoothness condition that ties post-strata together.

3.   STAGE 1: ESTIMATING Π

There are many ways to estimate the functions in Π, including local logistic multinomial regression and other parametric regression models. We use a nonparametric conditional density estimator by Hall et al. (2004). Let $m_i = 1$ if the ith unit is captured at least once, and $m_i = 0$ otherwise. Assume that each m_i is the outcome of a Bernoulli variable M_i. From (2), we have $\pi(y, x_{i\cdot}) = P(Y_{i\cdot} = y | M_i = 1, x_{i\cdot})$ for each $y \neq \vec{0}$. Suppose that each vector $x_{i\cdot}$ is a realization of some random variable X. Let $f_M(x_{i\cdot}) := P(X = x_{i\cdot} | M_i = 1)$ and $g_M(y, x) := \pi(y, x) f_M(x)$. Then,

$$g_M(y_{i\cdot}, x_{i\cdot}) = P(Y_{i\cdot} = y_{i\cdot} | X = x_{i\cdot}, M_i = 1) P(X = x_{i\cdot} | M_i = 1)$$
$$= P(y_{i\cdot}, x_{i\cdot} | M_i = 1).$$

Each of g_M and f_M can be estimated directly from the observable data (i.e., units with $m_i = 1$), and the conditional density of any nonzero capture pattern y given $X = x$ is

$$\pi(y, x) = \frac{g_M(y, x)}{f_M(x)}.$$

A nonparametric estimator of g_M uses smoothing parameters for both x and for the multinomial outcome y. However, we set the y bandwidth to zero (no smoothing), since Stage 2 of our two-stage process smooths over y in a more comprehensive way (see Section 4). With no smoothing over y, we are interested in only the normalized vectors of weights $w^i = (w_1^i, \cdots, w_{n_c}^i)$ such that the fitted values for the functions in Π are

$$\hat{\pi}(y, x_{i\cdot}) = \sum_{t=1}^{n_c} w_t^i I(y = y_{t\cdot}) \ (i = 1, ..., n_c; \ y \neq \vec{0}). \tag{5}$$

For example, if $\hat{\Pi}$ is a Gaussian smoother, and $f^i(\cdot, D)$ is the multivariate Gaussian density centered at $x_{i\cdot}$ with a covariance matrix D of bandwidth parameters, then $w_t^i = f^i(x_{t\cdot}, D) / \sum_t f^i(x_{t\cdot}, D) \ (i = 1, ..., n_c; \ t = 1, ..., n_c)$.

The array of estimated local cross-classification rates is $\widehat{\Pi}_i := \{\hat{\pi}(y, x_{i\cdot}) : y \neq \vec{0}\}$. Let $a(y)$ denote the array of indicators $\{I(z = y) : z \in \mathcal{Y}_k\}$. By (5), $\widehat{\Pi}_i = \sum_{t=1}^{n_c} w_t^i a(y_{t\cdot})$, a weighted-average array which gives greatest weight to observations with covariates close to $x_{i\cdot}$. For additional intuition about this notation, compare $\widehat{\Pi}_i$ with the cross-classification $c = \sum_{i=1}^n a(y_{i\cdot})$. In particular, if all the weights were 1, then each $\widehat{\Pi}_i$ would be the same as the observable part of c.

4.   STAGE 2: IMPUTING $\pi(\vec{0}, x)$

At a high level, Stage 2 involves fitting a log-linear model $\mathcal{M}_i$ to the local cross-classification $\widehat{\Pi}_i$, and then using the fitted model to project a value for the missing cell $\pi(\vec{0}, x_{i\cdot})$ $(i = 1, ..., n_c)$. Our framework permits maximal generality by allowing a separate model to be selected for every distinct point that is observed in the covariate space.

Stage 2 builds on the result of Stage 1, taking as given the vectors of smoothing weights w^i $(i = 1, ..., n_c)$. Hence, a mixture of multinomials $\sum_{t=1}^{n_c} w_t^i a(Y_{t\cdot})$ with conditional probabili-

ties $pr\{a(Y_{i\cdot}) = a(y_{i\cdot})|M_i = 1\} = \pi(y_{i\cdot}, x_{i\cdot})$ is assumed to generate $\widehat{\Pi}_i$ $(i = 1, ..., n_c)$. We now discuss the details of selecting and fitting a log-linear model for $\widehat{\Pi}_i$.

4·1. *Short review of log-linear models*

Log-linear models provide flexible ways to represent the cross-classification c as the outcome of a multinomial random variable. Classical log-linear models formally assume homogeneity, which says that $p(i_1, y) = p(i_2, y) =: p(y)$ for all $i_1, i_2 \in \{1, ..., n\}$. Independence between units is also assumed, so that the array of counts c is a realization of a multinomial random variable with n trials from the probability array $\{p(y)\}_{y \in \mathcal{Y}_k}$. Given a vector of parameters $u = (u_0, u_1, u_2, u_{12})$, a simple log-linear model is

$$\log p(y; u) = u_0 + u_1 y_1 + u_2 y_2 + u_{12} y_1 y_2 \; (y \in \mathcal{Y}_k), \tag{6}$$

where y_j denotes the jth element of the vector y $(j = 1, ..., k)$. The parameters u_1, u_2 are called list effects, and u_{12} represents the interaction between the first and second list. If there are more than two lists, additional parameters may be included to describe the other list interactions. For example, with k lists, the highest-order list interaction is denoted $u_{1\cdots k}$.

Parameter estimates can be found by maximizing the multinomial conditional likelihood

$$L_c(u|c \setminus c_0) = \frac{n_c!}{\prod_{y \neq \vec{0}} c_y!} \prod_{y \neq \vec{0}} \pi(y; u)^{c_y}, \tag{7}$$

where $\pi(y; u) := p(y; u)/(1 - p(\vec{0}; u))$ (Sanathanan, 1972; Fienberg, 1972). Given a maximum likelihood estimate $p(\vec{0}; \hat{u})$, the marginal likelihood

$$\frac{n!}{n_c!(n - n_c)!} p(\vec{0}; \hat{u})^{n - n_c} \{1 - p(\vec{0}; \hat{u})\}^{n_c}$$

is maximized over n to obtain a population estimate $\hat{n}$.

The cross-classification c has only $2^k - 1$ observable cells, and a unique maximizer of (7) exists for a model with at most $2^k - 1$ parameters. Thus, a model with exactly $2^k - 1$ parameters is called saturated, providing a perfect fit for the observed relative frequencies in the cells of c. With only two lists, one may take $u_{12} := 0$ in (6) to get a saturated hierarchical log-linear model, and maximizing the conditional and marginal likelihoods gives the Petersen estimator for the missing cell, $\hat{c}_{00} := \hat{n} - n_c = c_{10} c_{01}/c_{11}$.

4·2. *Local log-linear models*

Recalling (2), the conditional multinomial probabilities may vary as a function of some vector of unit-level covariates such as age or size. The Petersen estimator can be modified trivially to get the following unsurprising result: If $\pi(y, x)$ is constant in x, then the conditional maximum likelihood estimate of $\pi(\vec{0}, x) = \pi\{(0, 0), x\}$ is

$$\hat{\pi}\{(0, 0), x\} := \frac{\hat{\pi}\{(1, 0), x\}\hat{\pi}\{(0, 1), x\}}{\hat{\pi}\{(1, 1), x\}}, \tag{8}$$

where $\hat{\pi}(y, x) := c_y/n_c$ $(y \neq \vec{0})$ is the direct estimate for each observable capture pattern.

When $\pi(y, x)$ is not constant in x, (8) may still give a consistent estimator. It is easiest to see this when x can take on only S different values. Then the data can be partitioned into S classes, or post-strata, resulting in S separate cross-classification arrays $\{c(s)\}_{s=1}^S$ such that $\sum_s c(s) = c$. The saturated log-linear model may be fitted separately on each post-stratum; and the estimate of the missing cell is then of the form $\hat{c}_{00} = \sum_s \hat{c}_{00}(s) = \sum_s c_{01}(s)c_{01}(s)/c_{11}(s)$. In particular,

(8) clearly applies if we take $\hat{\pi}_S(y, x) := c_y(s_x)/n_c(s_x)$ where s_x denotes the index of the post-stratum containing x, and $n_c(s)$ is the number of observed units in post-stratum s.

The variability of $\hat{\pi}_S$ tends to increase as the number S of post-strata grows, since fewer observations are in each post-stratum. We can control the variance by imposing a smoothness assumption across post-strata. For example, $\hat{\pi}_S$ can be replaced with a kernel regression or parametric regression estimator $\hat{\pi}$. Assuming that $\pi(y, x)$ is sufficiently smooth to allow consistency of an estimator $\hat{\pi}(y, x)$, it follows that $\hat{\pi}(\vec{0}, x)$ as in (8) is consistent. Hence, we localize the u-terms by allowing them to take a different value for each covariate vector x, and (6) becomes

$$\log \pi\{y; u(x)\} = u_0(x) + u_1(x)y_1 + u_2(x)y_2 \ (y \in \mathcal{Y}_2)$$

for two lists. Similarly, for three lists, a saturated local log-linear model is

$$\log \pi\{y; u(x)\} = u_0(x) + u_1(x)y_1 + u_2(x)y_2 + u_3(x)y_3 \\ + u_{12}(x)y_1y_2 + u_{13}(x)y_1y_3 + u_{23}(x)y_2y_3 \ (y \in \mathcal{Y}_3). \tag{9}$$

Various models are obtained as submodels of (9) by removing terms. For example, the independence model for three lists leaves out the interaction terms to encode the assumption that the probability of capture on each list is independent of the event of capture on any other list:

$$\log \pi\{y; u(x)\} = u_0(x) + u_1(x)y_1 + u_2(x)y_2 + u_3(x)y_3. \tag{10}$$

4·3. *Estimating local log-linear parameters*

The likelihood of local log-linear model parameters $u(x)$ with respect to the smoothed data $\widehat{\Pi}_1, ..., \widehat{\Pi}_{n_c}$ is

$$\prod_{i=1}^{n_c} pr\left\{ \sum_{t=1}^{n_c} w_t^i a(Y_{t\cdot}) = \widehat{\Pi}_i \middle| u(x) \right\}.$$

This product is maximized when each individual term is maximized. Regarding each term of the product as a local conditional likelihood given the data $\widehat{\Pi}_i$, let

$$L_i = L\{u(x_{i\cdot})|\widehat{\Pi}_i\} := pr\left\{ \sum_{t=1}^{n_c} w_t^i a(Y_{t\cdot}) = \widehat{\Pi}_i \middle| u(x_{i\cdot}) \right\} \ (i = 1, ..., n_c).$$

In the special case for which the weights w^i come from a boxcar kernel and the capture pattern is homogeneous over the support of the kernel, L_i can be written in a form similar to (7) and analyzed by standard methods. Otherwise, exact evaluation of L_i is not straightforward. We approximate L_i by pretending that the mixture $\widehat{\Pi}_i$ is in fact multinomial, as follows.

For each $i \in \{1, ..., n_c\}$ and nonzero $y \in \mathcal{Y}_k$, pick a positive number λ_i and define $\hat{\pi}^*(y, x_{i\cdot})$ to be the unique value such that $\lambda_i\hat{\pi}^*(y, x_{i\cdot})$ is equal to $\lambda_i\hat{\pi}(y, x_{i\cdot})$ rounded to the nearest integer. We pick λ_i to be large, say $\lambda_i = 10^4$, so that the difference $\{\hat{\pi}^*(y, x_{i\cdot}) - \pi(y, x_{i\cdot})\}$ is negligible. The result is an integer-valued cross-classification array $\widehat{\Pi}_i^* = [\lambda_i\hat{\pi}^*(y, x_{i\cdot}) : y \neq \vec{0}]$ with essentially the same proportionality structure as $\widehat{\Pi}_i$, so that $\widehat{\Pi}_i^* \approx \lambda_i\widehat{\Pi}_i$. Since the element-wise differences $(\widehat{\Pi}_i^* - \lambda_i\widehat{\Pi}_i)$ are negligible, any log-linear model for $\widehat{\Pi}_i^*$ must fit $\widehat{\Pi}_i$ equally well in terms of any scale-free measure of goodness of fit. Therefore, we fit a log-linear model to $\widehat{\Pi}_i^*$ by assuming that $\widehat{\Pi}_i^*$ is a draw from a multinomial distribution with λ_i trials, with the likelihood function

$$L_c^*\{u(x_{i\cdot})|\widehat{\Pi}_i^*\} = \frac{\lambda_i!}{\prod_{y\neq\vec{0}}\{\lambda_i\hat{\pi}^*(y, x_{i\cdot})\}!} \prod_{y\neq\vec{0}} \pi\{y; u(x_{i\cdot})\}^{\lambda_i\hat{\pi}^*(y,x_{i\cdot})}. \tag{11}$$

Parameter estimates $\hat{u}(x_{i\cdot})$ are obtained by maximizing (11) subject to the constraint that the multinomial probabilities sum to 1, resulting in the estimate $\hat{\pi}(\vec{0}, x_{i\cdot}) := \pi\{\vec{0}, \hat{u}(x_{i\cdot})\}$ of the unknown $\pi(\vec{0}, x_{i\cdot})$. We refer to this method as pseudo-multinomial maximum likelihood estimation.

Estimation by pseudo-multinomial maximum likelihood is essentially a sliding-window version of post-stratification if the weights w^i correspond to a boxcar-like kernel. The traditional method, fitting log-linear models to disjoint post-strata, and our new approach, pseudo-multinomial maximum likelihood on overlapping kernels, may both incur bias if $\pi(y, x)$ is particularly variable with respect to x over the support of any given post-stratum or kernel evaluated at a fixed $x_{i\cdot}$. Also, population estimates have high variance if the post-strata are too small or if the smooth post-stratifier $\widehat{\Pi}$ undersmooths.

The fact that parameter estimation is done separately for each population unit allows us select a separate model at each x. A more rigorous notation for $u(x)$ in (11) could be something like $u\{\mathcal{M}(x)\}$ to reflect that u is the parameter vector corresponding to whichever model $\mathcal{M}$ is selected at x. Specifically, for distinct points in the covariate space x_1 and x_2, parameter vectors $u(x_1)$ and $u(x_2)$ may differ in length. This expanded notation is omitted for brevity.

4·4. *Local Model Selection*

Several old strategies for log-linear model selection exist (Benedetti & Brown, 1978). Fienberg (1972) recommended a nuanced and somewhat ad hoc method centered around likelihood ratio tests, and stepwise regression based on an information criterion has seen recent use in capture-recapture applications (Aaron et al., 2003; Murphy, 2009). Each of these approaches can be applied for the local log-linear model selection problem with appropriate modifications.

We choose to implement the Schwarz information criterion due to its simplicity and convenience. With a minor modification, this criterion is applicable to the local model selection problem in conjunction with pseudo-multinomial maximum likelihood estimation. The reason for a modification is that the apparent number of degrees of freedom in a log-linear fit of $\widehat{\Pi}_i^*$ has no connection to reality after the scaling by the arbitrarily chosen λ_i. Specifically, a large inflation factor λ_i implies a large total cell count in $\widehat{\Pi}_i^*$, tending to cause overfitting.

We elaborate on this point. Using the pseudo-multinomial likelihood (11), the standard Schwarz information criterion objective function to be minimized is $-2 \log L_i^* + q \log(\lambda_i)$, where q is the number of log-linear parameters. Increasing λ_i means that there are more terms in the product L_i^*, so $-2 \log L_i$ grows at least linearly in λ_i, while $q \log(\lambda_i)$ grows only log-linearly. Therefore, when λ_i is large, $-2 \log L_i^*$ tends to be more important than $q \log(\lambda_i)$, leading to the selection of a log-linear model with too many parameters.

Ideally, the number of terms in the pseudo-multinomial likelihood to be used for the information criterion should reflect the number of observed population units underlying the locally averaged multinomial frequencies $\widehat{\Pi}_i$. A reasonable proxy for this quantity, which we will call the local degrees of freedom, is the sum of the conditional regression weights for the ith unit: $n_i := \sum_{t=1}^{n_c} w_t^i$. Since the sum of the elements of the array $(n_i/\lambda_i)\widehat{\Pi}_i^*$ is n_i, we replace the standard Schwarz information criterion objective function with $-2(n_i/\lambda_i) \log(L_i^*) + q \log(n_i)$. A similar modification could be used with the Akaike information criterion.

5. Variance estimation

The multi-stage structure of our method makes it difficult to compute the variance directly. Instead, we simulate the sampling distribution of our estimator using a method that is similar to the parametric bootstrap described by Zwane & van der Heijden (2003).

Let $(x^c_{..}, \cdot)$ denote the covariates of the n_c observed units. The fraction $1/\hat{\psi}(x_{i.})$ in (1) is the number of units that are represented by the ith unit. That is, the contribution of the ith unit in the Horvitz-Thompson sum $\hat{n}$ can be decomposed as the sum of 1, representing the ith unit, and $o_i := 1/\hat{\psi}(x_{i.}) - 1$ additional units that were not captured. Let o_i^{int} and o_i^{dec} denote the integer and fractional components of o_i, respectively, such that $o_i = o_i^{int} + o_i^{dec}$, where all quantities are non-negative. Let o'_i denote the sum of o_i^{int} with the outcome of a Bernoulli random variable with success probability o_i^{dec}.

For each $i \in \{1, ..., n_c\}$, we insert o'_i new units with the covariate $x_{i.}$. With the insertions, the set of covariates represents the population that is assumed by our model, and this set is denoted as $(x_{..}, \cdot)^{sim}$. Replacing o_i with its random perturbation o'_i introduces an element of randomness that is not included in our model, and this should slightly inflate the variance of our simulation outcomes, leading to conservative confidence intervals.

Finally, the capture pattern for the ith individual $y_{i.}$ is simulated as a multinomial random variable with unit-level multinomial probabilities $\hat{r}(y_{i.}, x_{i.}) := \hat{\psi}(x_{i.})\hat{\pi}(y_{i.}, x_{i.})$. The result is a simulated population $(x_{..}, y_{..})^{sim}$ with covariates and capture patterns observed for all units. Deleting all units with capture pattern $\vec{0}$ gives the simulated data $(x^c_{..}, y^c_{..})^{sim}$, and applying a smooth post-stratification methods gives an estimate $\hat{c}_0^{sim}$ of the number of units not observed. Replicating this procedure leads to bootstrap confidence intervals.

6. SEVERAL IMPORTANT LOCAL LOG-LINEAR MODELS

Our adaptation of the Schwarz information criterion for model selection in Section 4·4 is intended as a first-generation solution to the local model selection problem. In this section, we take a small step backwards to discuss several specific log-linear models that are interesting due to interpretability or historical significance in the capture-recapture literature.

The independence model (10) can be further constrained by requiring that the list effects are all equal to a single parameter, $u_\Sigma := u_1 = \cdots = u_k$. This leads to a local model of independence with equal catch-ability across lists, an extremely sparse model with only one free parameter:

$$\log \pi\{y; u(x)\} = u(x) + u_\Sigma(x) \sum_{j=1}^{k} y_j. \tag{12}$$

For basic local log-linear models with no highest-order interaction term (i.e., $u_{1...k}(x) := 0$) such as (9), (10), and (12), eliminating all of the explicit $u(x)$-terms from the model equations leads to

$$\pi(\vec{0}; u(x)) = \frac{\prod_{y \in O} \pi(y; u(x))}{\prod_{z \in E} \pi(z; u(x))}, \tag{13}$$

where O is the set of capture patterns with entries summing to an odd number, and E is the set of nonzero capture patterns summing to an even number (Fienberg, 1972). The saturated model (9) is particularly convenient to implement in conjunction with (13) because the model always fits the observable data exactly. For example, Zwane & van der Heijden (2004) directly estimated $\pi(\vec{0}, x)$ by substituting the Stage-1 function estimates $\hat{\pi}(y, x)$ in place of $\pi(y; u(x))$ in the right-hand side of formula (13) without giving attention to the log-linear equations.

The saturated model (9) is appealing in its convenience and flexibility, but the resulting estimates can be extremely unstable due to overfitting. We propose a way to stabilize the estimates by averaging the saturated model (9) with the two-parameter model (12), as follows. Let $\Pi_i(\hat{u}) = [\pi\{y; \hat{u}(x_{i.})\} : y \neq \vec{0}]$ denote the conditional multinomial probabilities implied by

model (12) given parameter estimates $\hat{u} = (\hat{u}_0, \hat{u}_\Sigma)$. Let $\nu_i = min_{y \neq 0}\hat{\pi}(y, x_{i\cdot})$. Recalling that n_i denotes the effective degrees of freedom for $\widehat{\Pi}_i$, define a mixing constant $\alpha_i \in (0, 1)$ as

$$\alpha_i = \frac{n_i \nu_i}{1 + n_i \nu_i},$$

and define a weighted average $\Pi_i(\hat{u}, \alpha) := (1 - \alpha_i)\Pi_i(\hat{u}) + \alpha_i \widehat{\Pi}_i$ $(i = 1, ..., n_c)$, where the linear combination of arrays is evaluated element-wise. Plugging $\Pi_i(\hat{u}, \alpha)$ into the right-hand side of (13) gives an estimate for $\pi(\vec{0}; u(x))$. Heuristically, the constant α_i is small, putting greater weight on the sparse fit $\Pi_i(\hat{u})$, when the effective degrees of freedom is not large enough to stabilize the smallest element of the saturated fit $\widehat{\Pi}_i$. We call this the adjusted saturated model.

Finally, we mention a non-hierarchical log-linear model with a single list-interaction parameter that was inspired by the Rasch model for educational testing. The quasi-symmetry model is derived from log-normal unit-level random effects and takes the form

$$\log \pi(y; u) = u + u_1 y_1 + u_2 y_2 + u_3 y_3 + u_\Sigma \left(\sum_{j=1}^{k} y_j \right)^2, \qquad (14)$$

with moment restrictions that constrain u_Σ to be greater than zero (Darroch et al., 1993). The random effects interpretation of the quasi-symmetry model has been invoked to model populations with heterogeneous capture probabilities without explicitly controlling for observable covariates x. Supposing that the covariates x do not fully explain the heterogeneity in a population, one could apply the quasi-symmetry model as a local log-linear model (modifying the parameters to become functions of x as in previous examples) to model any unexplained heterogeneity at each level of x. However, prior to trusting the results of any latent-class or random-effects model for heterogeneity, we strongly recommend careful consideration of the non-identifiability studies of Link (2003a) and Mao (2008).

7. A Simple Application

7·1. *How Many Birds Species Can be Observed in the U.S. and Canada?*

We estimate the number of bird species using the North American Breeding Bird Survey for continental North America north of Mexico (Sauer et al., 2011). Table 2 displays c, the cross-classification of species observed in the years 2009 - 2011, treating each year as a separate list. For example, exactly 581 species were observed in all three years, and 18 species were observed only in 2009.

Table 2. *Cross-classification of species observed over three years*

		In 2011	Not in 2011
In 2010	In 2009	581	13
	Not in 2009	10	10
Not in 2010	In 2009	11	18
	Not in 2009	21	c_0

Define a covariate x as the reverse of the rank ordering of the observed species based on the total number of times that each species was observed. For example, the species that was observed most often over the three years has covariate $x = 664$, as 664 distinct species were observed. The obvious interpretation of x is that species with a high value of x are easy to observe. Compared to

covariates uses previously to model heterogeneity in the detectability of birds, such as wingspan, our covariate appears to be a relatively direct proxy for species detectability.

We estimate the conditional probability functions Π using the np package (Hayfield & Racine, 2008) in the R statistical software (R Core Team, 2012). The estimated functions $\widehat{\Pi}$, the result of Stage 1, appear in each panel of Figure 1 in a stacked form. These seven curves sum to the horizontal line at height 1, labeled "010", reflecting the identity that the conditional multinomial capture pattern probabilities must sum to 1 at each x. We subsequently impute $\pi(0, x_i)$ by four different methods, plotting the results as 664 points in each of the panels of Figure 1.

The estimates for the independence model (10) were obtained using pseudo-multinomial maximum likelihood. The result is displayed as the top curve in panel (a), labeled "000". For example, near $x = 1$ the distance between the top curve and the horizontal line below it is nearly 0·5. This indicates that the independence model imputes nearly 0·5 unobserved units corresponding to each observed unit. For all units with $x \geq 100$, the independence model imputes approximately zero unobserved units, as the top curve is nearly coincident with the next-to-top curve.

The independence model is valid if the event that a species is observed in one year is independent of the event that this species is observed in another year. However, a positive dependence between years is plausible if the experience that a bird watcher gains in sighting a certain rare species in one year increases the probability of similar sightings in years following. Therefore, it may be appropriate to include interaction terms.

The adjusted saturated model in panel (b) was described in Section 6. Applying the saturated model (13) directly is not practical because some of the denominator terms approach zero for $x > 150$, causing the formula to become highly unstable or undefined.

Panel (c) in Figure 1 shows the result of applying a stepwise local log-linear model search using the Schwarz information criterion. The discontinuities in the imputation curve (labeled "000") mark the points at which the choice of local log-linear model changed. One could easily smooth across adjacent models to remove the discontinuities, using something like local linear regression. A more-fundamental issue is that the selected models are extremely sparse. For example, the short horizontal section of the imputation curve corresponds to the 0-free-parameter model, which has only an intercept term, assigning equal weight (1/7) to all capture patterns.

The small sizes of the models preferred by the Schwarz information criterion reflect the unfortunate reality that the effective degrees of freedom is small, on the order of 50, providing limited information for a multinomial with seven outcomes. A natural reaction is to try over-smoothing the conditional density estimate $\widehat{\Pi}$ to increase the degrees of freedom. Panel (d) shows the outcome of over-smoothing, with the regression bandwidth (in x) increased to 250 from the cross-validation-selected bandwidth of only 27. The estimate of 724 missing species is implausibly high, although it is encouraging that the model imputed no missing species for $x > 400$. Excessive over-smoothing clearly defeats the purpose of local model fitting; choosing an optimal degree of over-smoothing in this context remains an open problem.

As a point of reference from the existing capture-recapture literature, we fit the quasi-symmetry model (14). Following Darroch et al. (1993), we ignore the restriction $u_\Sigma > 0$ during parameter estimation (nevertheless, we obtained $\hat{u}_\Sigma > 0$ for the point estimate), resulting in an implausibly large estimate $\hat{c}_0 = 1744$. An alternative is to apply the quasi-symmetry model locally, since regression smoothing and latent covariates may leave significant unexplained heterogeneity at each level of x. Applying the quasi-symmetry model locally is not straightforward in this application due to numeric instability in the estimates where $x > 150$ (here, most of the functions in $\widehat{\Pi}$ approach 0). However, if we assume that a negligible number of species is missing for $x > 150$ and consider imputed values only for $x < 150$, we arrive at a more reasonable point
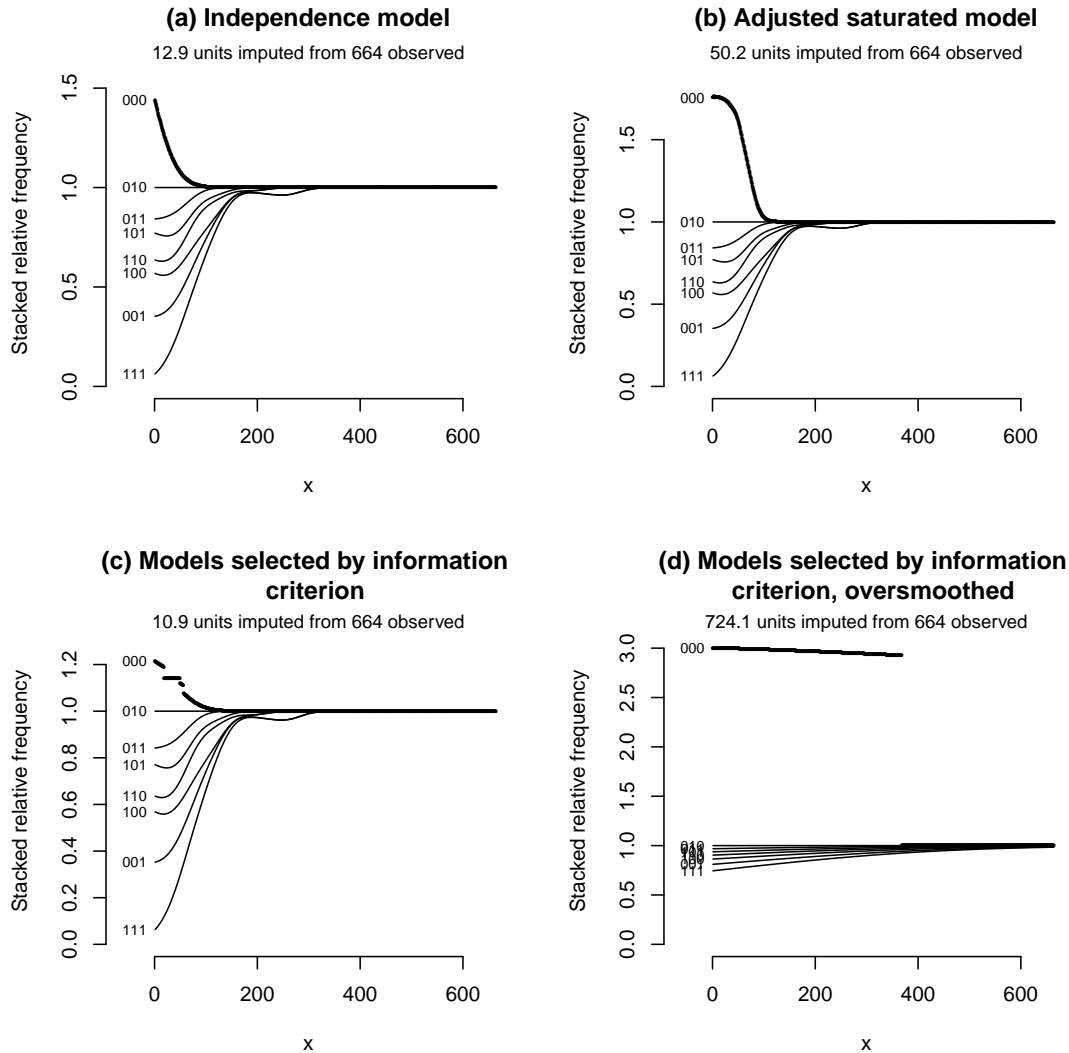
(a) Independence model

12.9 units imputed from 664 observed

(b) Adjusted saturated model

50.2 units imputed from 664 observed

(c) Models selected by information criterion

10.9 units imputed from 664 observed

(d) Models selected by information criterion, oversmoothed

724.1 units imputed from 664 observed

Fig. 1. The four panels each display the result of a separate estimation method. In each panel, the horizontal axis is the species rank x. The relative frequencies of the capture patterns (i.e., "111", "001", ...) are plotted as functions of x in a stacked form. For example, the curve labeled "001" represents the sum $\pi\{(1,1,1); \hat{u}(x)\} + \pi\{(0,0,1); \hat{u}(x)\}$. The relative frequencies of observable capture patterns sum to 1, the horizontal line, labeled "010". Above the horizontal line, the imputed values are plotted as $\hat{\pi}\{(0,0,0), x_i\} + 1$ $(i = 1, ..., 664)$. The seven relative frequency curves plotted in panel (d) are much straighter than they appear in the other panels due to intentional over-smoothing in the Stage-1 smoothing process.

estimate of $\hat{c}_0 = 85$. Broadly applying this model may require mixing with a simpler model as in the adjusted saturated model, and we save this for future work.

Bootstrapped 90% confidence intervals and standard errors for the number of unobserved species are summarized for several models in Table 3.

Table 3. *Bootstrapped estimates of variability of $\hat{c}_0$ in several models*

Model	$se(\hat{c}_0)$	90% Confidence Interval for c_0
Independence	3·8	(4·3, 16)
Adjusted Saturated	48	(42, 181)
Schwarz information criterion	4·4	(7·8, 16)
Quazi-symmetry (non-local)	1940	(640, 4900)

The application of capture-recapture methods to three years of data raises the obvious question: Why not extend the model to incorporate all available years of data? Indeed, the Breeding Bird Survey data goes back to 1965. However, the assumption of a closed population may fail over long spans of time, as certain species go extinct, and new species evolve or change their geographic region of preference. Effective population size estimation on a 3-year moving window could, in principle, reveal changes in species richness over time. A separate consideration is that not using data earlier than 2009 allows us to use the previous years of data as a partial validation of our method. The data collection format was standardized in 1997; the data from 1997 to 2011 reveals 704 distinct species, strongly suggesting that the independence model (a) and Schwarz information criterion model (c) are suboptimal in their current form.

In applying the Horvitz-Thompson estimator (1) or (4), we ignored the point made by Alho (1990) that the estimator is not consistent if the detection probability $\psi(x)$ approaches 0, even if $\psi(x)$ were known. In application, $\psi(x)$ must be estimated, and the necessary condition is arguably much more strict: $\psi(x)$ must be substantially greater than zero if an estimate $\hat{\pi}(\vec{0}, x)$ is to have a useful level of precision. Specifically, the detection probabilities in the left tail of the distribution of x in Figure 1 may be too low to estimate accurately. This point deserves more attention in many capture-recapture studies, including previous studies using Breeding Bird Survey data such as Boulinier et al. (1998) and Dorazio & Royle (2003).

8. DISCUSSION

The capture-recapture problem is fundamentally a missing data problem. The missing quantity of interest is n, the population size, but, perhaps more relevantly, the covariate values x_i. are missing for $i > n_c$. The nature of this missingness is arguably of the worst possible kind, because it is reasonable to suppose that the units which are not observed are not observed precisely because they are different from the observed units, not only in the distribution of covariates but also in how capture probabilities depend on covariates. Differences between the training data and the test data plague every prediction problem, but it is not generally the case that the prediction data is different from the training data by default. This difference sets apart capture-recapture as an exceptionally difficult and risky estimation task.

Controlling for covariates is effective only if the observable covariates explain much of the heterogeneity in capture probabilities, and this condition is not always attainable. Regardless, at least attempting to use available covariates is a necessary first step to understanding heterogeneity. The necessity of modeling heterogeneity on covariates has been controversial. Much of the capture-recapture literature attempts to incorporate heterogeneity effects without using covariates (Pledger & Phillpot, 2008). However, Link showed that the population size n is often not identifiable across alternative heterogeneity models (Link, 2003a). Pledger countered that

model misspecification is a relatively minor concern if several different kinds of models all lead to similar estimates (Pledger, 2005), a position that was rejected by Link (2003b).

Smooth post-stratification points to several avenues of future work. One key question stems from our two-stage approach. The first stage is estimating the functions in Π, and this involves a variable selection and/or bandwidth selection problem. The second bias-variance tradeoff occurs in the local selection of log-linear models for imputing $\pi(0, x)$, which may emphasize parsimony to a greater or lesser degree. Each of these stages has its own bias-variance tradeoff. Currently, the two trade-offs are optimized separately, perhaps by using cross-validation in the the first stage and by some information criterion in the second stage. Aesthetically, and perhaps more substantively, it is desirable to unify these two modeling problems.

Local log-linear models enable a high degree of specificity; model selection (not only model fitting) may be performed separately for each observed population unit. While the flexibility to select a separate model for each population unit raises the specter of overfitting, the smoothed nature of $\widehat{\Pi}$ ensures that the models are highly correlated across units. Therefore, it is unclear how to count degrees of freedom globally for the proposed multi-stage estimation, and a penalty on the flexibility of the model selection procedure may be appropriate.

REFERENCES

AARON, D. J., CHANG, Y.-F., MARKOVIC, N. & LaPORTE, R. E. (2003). Estimating the lesbian population: A capture-recapture approach. *Journal of Epidemiology and Community Health* **57**, 207–209.

ALHO, J. M. (1990). Logistic regression in capture-recapture models. *Biometrics* **46**, 623–635.

BENEDETTI, J. K. & BROWN, M. B. (1978). Strategies for the selection of log-linear models. *Biometrics* **34**, 680–686.

BOULINIER, T., NICHOLS, J. D., SAUER, J. R., HINES, J. E. & POLLOCK, K. H. (1998). Estimating species richness: The importance of heterogeneity in species detectability. *Ecology* **79**, 1018–1028.

CHEN, S. X. & LLOYD, C. J. (2002). Estimation of population size from biased samples using non-parametric binary regression. *Statistica Sinica* **12**, 505–518.

CHEN, S. X., TANG, C. Y. & VINCENT T. MULE, J. (2010). Local post-stratification in dual system accuracy and coverage evaluation for the U.S. Census. *Journal of the American Statistical Association* **105**, 105–119.

DARROCH, J. N., FIENBERG, S. E., GLONEK, G. F. V. & JUNKER, B. W. (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *Journal of the American Statistical Association* **88**, 1137–1148.

DORAZIO, R. M. & ROYLE, J. A. (2003). Mixture models for estimating the size of a closed population when capture rates vary among individuals. *Biometrics* **59**, 351–364.

FIENBERG, S. E. (1972). The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika* **59**, 591.

FIENBERG, S. E., JOHNSON, M. S. & JUNKER, B. W. (1999). Classical multilevel and Bayesian approaches to population size estimation using multiple lists. *Journal of the Royal Statistical Society* **162**, 383–405.

HALL, P., RACINE, J. & LI, Q. (2004). Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association* **99**, 1015–1026.

HAYFIELD, T. & RACINE, J. S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software* **27**.

HUGGINS, R. (1989). On the statistical analysis of capture experiments. *Biometrika* **76**, 133–140.

HWANG, W.-H. & HUGGINS, R. (2011). A semiparametric model for a functional behavioural response to capture in capture-recapture experiments. *Australian & New Zealand Journal of Statistics* **53**, 191202.

LINK, W. A. (2003a). Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics* **59**, 1123–1130.

LINK, W. A. (2003b). Reply to a paper by Holzmann, Munk, and Zucchini. *Biometrics* **59**, 1123–1130.

MAO, C. X. (2008). On the nonidentifiability of population sizes. *Biometrics* **64**, 977–981.

MURPHY, J. (2009). Estimating the World Trade Center tower population on September 11, 2001: A capture-recapture approach. *American Journal of Public Health* **99**, 65–67.

ODUM, E. P. & PONTIN, A. J. (1961). Population density of the underground ant, lasius flavus, as determined by tagging with p32. *Ecology* **42**, 186–188.

PLEDGER, S. (2005). The performance of mixture models in heterogeneous closed population capture-recapture. *Biometrics* **61**, 868–876.

PLEDGER, S. & PHILLPOT, P. (2008). Using mixtures to model heterogeneity in ecological capture-recapture studies. *Biometrical Journal* **50**, 1022–1034.

POLLOCK, K. H. (1976). Building models of capture-recapture experiments. *Journal of the Royal Statistical Society* **25**, 253–259.

POLLOCK, K. H., HINES, J. E. & NICHOLS, J. D. (1984). The use of auxiliary variables in capture-recapture and removal experiments. *Biometrics* **40**, 329–340.

R CORE TEAM (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

RUNESON, P. & WOHLIN, C. (1998). An experimental evaluation of an experience-based capture-recapture method in software code inspections. *Empirical Software Engineering* **3**, 381–406.

SANATHANAN, L. (1972). Estimating the size of a multinomial population. *Annals of Mathematical Statistics* **43**, 142–152.

SAUER, J. R., HINES, J. E., FALLON, J. E., PARDIECK, K. L., D. J. ZIOLKOWSKI, J. & LINK, W. A. (2011). The North American Breeding Bird Survey, Results and Analysis 1966 - 2010. Version 12.07.2011 USGS Patuxent Wildlife Research Center, Laurel, MD.

STOKLOSA, J. & HUGGINS, R. M. (2012). A robust P-spline approach to closed population capture-recapture models with time dependence and heterogeneity. *Computational Statistics & Data Analysis* **56**, 408 – 417.

YIP, P. S. F., WAN, E. C. Y. & CHAN, K. S. (2001). A unified approach for estimating population size in capture-recapture studies with arbitrary removals. *Journal of Agricultural, Biological, and Environmental Statistics* **6**, 183–194.

ZWANE, E. & VAN DER HEIJDEN, P. (2003). Implementing the parametric bootstrap in capture-recapture models with continuous covariates. *Statistics & Probability Letters* **65**, 121–125.

ZWANE, E. & VAN DER HEIJDEN, P. (2004). Semiparametric models for capture-recapture studies with covariates. *Computational Statistics & Data Analysis* **47**, 729743.